



# Genome-Wide Diet-Gene Interaction Analyses for Risk of Colorectal Cancer

## Citation

Figueiredo, J. C., L. Hsu, C. M. Hutter, Y. Lin, P. T. Campbell, J. A. Baron, S. I. Berndt, et al. 2014. "Genome-Wide Diet-Gene Interaction Analyses for Risk of Colorectal Cancer." PLoS Genetics 10 (4): e1004228. doi:10.1371/journal.pgen.1004228. <http://dx.doi.org/10.1371/journal.pgen.1004228>.

## Published Version

doi:10.1371/journal.pgen.1004228

## Permanent link

<http://nrs.harvard.edu/urn-3:HUL.InstRepos:12152846>

## Terms of Use

This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at <http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA>

## Share Your Story

The Harvard community has made this article openly available.  
Please share how this access benefits you. [Submit a story](#).

[Accessibility](#)



# Genome-Wide Diet-Gene Interaction Analyses for Risk of Colorectal Cancer

Jane C. Figueiredo<sup>1\*</sup>, Li Hsu<sup>2</sup>, Carolyn M. Hutter<sup>3</sup>, Yi Lin<sup>2</sup>, Peter T. Campbell<sup>4</sup>, John A. Baron<sup>5</sup>, Sonja I. Berndt<sup>6</sup>, Shuo Jiao<sup>2</sup>, Graham Casey<sup>1</sup>, Barbara Fortini<sup>1</sup>, Andrew T. Chan<sup>7,8</sup>, Michelle Cotterchio<sup>9</sup>, Mathieu Lemire<sup>10</sup>, Steven Gallinger<sup>11</sup>, Tabitha A. Harrison<sup>2</sup>, Loic Le Marchand<sup>12</sup>, Polly A. Newcomb<sup>2</sup>, Martha L. Slattery<sup>13</sup>, Bette J. Caan<sup>14</sup>, Christopher S. Carlson<sup>2</sup>, Brent W. Zanke<sup>15</sup>, Stephanie A. Rosse<sup>2</sup>, Hermann Brenner<sup>16</sup>, Edward L. Giovannucci<sup>17</sup>, Kana Wu<sup>18</sup>, Jenny Chang-Claude<sup>19</sup>, Stephen J. Chanock<sup>6</sup>, Keith R. Curtis<sup>2</sup>, David Duggan<sup>20</sup>, Jian Gong<sup>2</sup>, Robert W. Haile<sup>21</sup>, Richard B. Hayes<sup>22</sup>, Michael Hoffmeister<sup>16</sup>, John L. Hopper<sup>23</sup>, Mark A. Jenkins<sup>23</sup>, Laurence N. Kolonel<sup>12</sup>, Conghui Qu<sup>2</sup>, Anja Rudolph<sup>19</sup>, Robert E. Schoen<sup>24</sup>, Fredrick R. Schumacher<sup>1</sup>, Daniela Seminara<sup>3</sup>, Deanna L. Stelling<sup>2</sup>, Stephen N. Thibodeau<sup>25</sup>, Mark Thornquist<sup>2</sup>, Greg S. Warnick<sup>2</sup>, Brian E. Henderson<sup>1</sup>, Cornelia M. Ulrich<sup>2,26</sup>, W. James Gauderman<sup>1</sup>, John D. Potter<sup>2,27</sup>, Emily White<sup>2</sup>, Ulrike Peters<sup>2\*</sup>, on behalf of CCFR, and GECCO<sup>1</sup>

**1** Keck School of Medicine, University of Southern California, Los Angeles, California, United States of America, **2** Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, Washington, United States of America, **3** Division of Cancer Control and Population Sciences, National Cancer Institute, Bethesda, Maryland, United States of America, **4** Epidemiology Research Program, American Cancer Society, Atlanta, Georgia, United States of America, **5** Department of Medicine, School of Medicine, University of North Carolina, Chapel Hill, North Carolina, United States of America, **6** Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, Maryland, United States of America, **7** Division of Gastroenterology, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts, **8** Channing Division of Network Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, United States of America, **9** Cancer Care Ontario, Toronto, Ontario, Canada, **10** Ontario Institute for Cancer Research, Toronto, Ontario, Canada, **11** Department of Surgery, University Health Network Toronto General Hospital, Toronto, Ontario, Canada, **12** Epidemiology Program, University of Hawaii Cancer Center, Honolulu, Hawaii, United States of America, **13** Department of Internal Medicine, University of Utah Health Sciences Center, Salt Lake City, Utah, United States of America, **14** Division of Research, Kaiser Permanente Medical Care Program, Oakland, California, United States of America, **15** Division of Hematology, Faculty of Medicine, The University of Ottawa, Ottawa, Ontario, Canada, **16** Division of Clinical Epidemiology and Aging Research, German Cancer Research Center, Heidelberg, Germany, **17** Channing Division of Network Medicine, Brigham and Women's Hospital, Boston, Massachusetts, United States of America, **18** Department of Nutrition, Harvard School of Public Health, Boston, Massachusetts, United States of America, **19** Division of Cancer Epidemiology, German Cancer Research Center, Heidelberg, Germany, **20** Translational Genomics Research Institute, Phoenix, Arizona, United States of America, **21** Stanford Cancer Institute, Palo Alto, California, United States of America, **22** Division of Epidemiology, New York University School of Medicine, New York, New York, United States of America, **23** Melbourne School of Population Health, The University of Melbourne, Melbourne, Australia, **24** Department of Medicine and Epidemiology, University of Pittsburgh Medical Center, Pittsburgh, Pennsylvania, United States of America, **25** Departments of Laboratory Medicine and Pathology and Laboratory Genetics, Mayo Clinic, Rochester, Minnesota, United States of America, **26** Division of Preventive Oncology, National Cancer Center for Tumor Diseases and German Cancer Research Center, Heidelberg, Germany, **27** Centre for Public Health Research, Massey University, Wellington, New Zealand

## Abstract

Dietary factors, including meat, fruits, vegetables and fiber, are associated with colorectal cancer; however, there is limited information as to whether these dietary factors interact with genetic variants to modify risk of colorectal cancer. We tested interactions between these dietary factors and approximately 2.7 million genetic variants for colorectal cancer risk among 9,287 cases and 9,117 controls from ten studies. We used logistic regression to investigate multiplicative gene-diet interactions, as well as our recently developed Cocktail method that involves a screening step based on marginal associations and gene-diet correlations and a testing step for multiplicative interactions, while correcting for multiple testing using weighted hypothesis testing. Per quartile increment in the intake of red and processed meat were associated with statistically significant increased risks of colorectal cancer and vegetable, fruit and fiber intake with lower risks. From the case-control analysis, we detected a significant interaction between rs4143094 (10p14/near *GATA3*) and processed meat consumption (OR = 1.17;  $p = 8.7 \times 10^{-9}$ ), which was consistently observed across studies ( $p$  heterogeneity = 0.78). The risk of colorectal cancer associated with processed meat was increased among individuals with the rs4143094-TG and -TT genotypes (OR = 1.20 and OR = 1.39, respectively) and null among those with the GG genotype (OR = 1.03). Our results identify a novel gene-diet interaction with processed meat for colorectal cancer, highlighting that diet may modify the effect of genetic variants on disease risk, which may have important implications for prevention.

**Citation:** Figueiredo JC, Hsu L, Hutter CM, Lin Y, Campbell PT, et al. (2014) Genome-Wide Diet-Gene Interaction Analyses for Risk of Colorectal Cancer. *PLoS Genet* 10(4): e1004228. doi:10.1371/journal.pgen.1004228

**Editor:** Christopher I. Amos, Dartmouth-Hitchcock Norris Cotton Cancer Center, United States of America

**Received:** September 12, 2013; **Accepted:** January 20, 2014; **Published:** April 17, 2014

This is an open-access article, free of all copyright, and may be freely reproduced, distributed, transmitted, modified, built upon, or otherwise used by anyone for any lawful purpose. The work is made available under the Creative Commons CC0 public domain dedication.

**Funding:** This work was supported by, GECCO: National Cancer Institute, National Institutes of Health, U.S. Department of Health and Human Services (U01 CA137088; R01 CA059045; R01 CA120582). CCFR: National Institutes of Health (RFA # CA-95-011) and through cooperative agreements with members of the Colon Cancer Family Registry and P.I.s. This genome wide scan was supported by the National Cancer Institute, National Institutes of Health by U01 CA122839. The content of this manuscript does not necessarily reflect the views or policies of the National Cancer Institute or any of the collaborating centers in the CFRs.

nor does mention of trade names, commercial products, or organizations imply endorsement by the US Government or the CFR. The following Colon CFR centers contributed data to this manuscript and were supported by National Institutes of Health: Australasian Colorectal Cancer Family Registry (U01 CA097735), Seattle Colorectal Cancer Family Registry (U01 CA074794) and Ontario Registry for Studies of Familial Colorectal Cancer (U01 CA074783). DACHS: German Research Council (Deutsche Forschungsgemeinschaft, BR 1704/6-1, BR 1704/6-3, BR 1704/6-4 and CH 117/-1), and the German Federal Ministry of Education and Research (01KH0404 and 01ER0814). DALs: National Institutes of Health (R01 CA48998 to MSL); HPFS was supported by the National Institutes of Health (P01 CA 055075, UM1 CA167552, R01 137178, and P50 CA 127003), and NHS by the National Institutes of Health (R01 137178, P01 CA 087969 and P50 CA 127003). OFCCR: National Institutes of Health, through funding allocated to the Ontario Registry for Studies of Familial Colorectal Cancer (U01 CA074783); see CCFR section above. Genetic analyses have been supported by a GL2 grant from the Ontario Research Fund, the Canadian Institutes of Health Research, the Cancer Risk Evaluation (CaRE) Program grant from the Canadian Cancer Society Research Institute and the Ontario Institute for Cancer Research, through generous support from the Ontario Ministry of Research and Innovation. PLCO: Intramural Research Program of the Division of Cancer Epidemiology and Genetics and supported by contracts from the Division of Cancer Prevention, National Cancer Institute, NIH, DHHS. Additionally, a subset of control samples were genotyped as part of the Cancer Genetic Markers of Susceptibility (CGEMS) Prostate Cancer GWAS (Yeager, M et al. Nat Genet 2007 May;39(5):645–9), Colon CGEMS pancreatic cancer scan (PanScan) (Amundadottir, L et al. Nat Genet. 2009 Sep;41(9):986–90 and Petersen, GM et al Nat Genet. 2010 Mar;42(3):224–8), and the Lung Cancer and Smoking study. The prostate and PanScan study datasets were accessed with appropriate approval through the dbGaP online resource (<http://cgems.cancer.gov/data/>) accession numbers 000207v1.p1 and phs000206.v3.p2, respectively, and the lung datasets were accessed from the dbGaP website (<http://www.ncbi.nlm.nih.gov/gap>) through accession number phs000093.v2.p2. Funding for the Lung Cancer and Smoking study was provided by National Institutes of Health (NIH), Genes, Environment and Health Initiative (GEI) Z01 CP 010200, NIH U01 HG004446, and NIH GEI U01 HG 004438. For the lung study, the GENEVA Coordinating Center provided assistance with genotype cleaning and general study coordination, and the Johns Hopkins University Center for Inherited Disease Research conducted genotyping. PMH: National Institutes of Health (R01 CA076366 to PAN). VITAL: National Institutes of Health (K05 CA154337). WHI: The WHI program was funded by the National Heart, Lung, and Blood Institute, National Institutes of Health, U.S. Department of Health and Human Services through contracts HHSN268201100046C, HHSN268201100001C, HHSN268201100002C, HHSN268201100003C, HHSN268201100004C, and HHSN271201100004C. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing Interests:** The authors have declared that no competing interests exist.

\* E-mail: janefigu@usc.edu (JCF); upeters@fhcr.org (UP)

† Membership of GECCO is provided in the Acknowledgments.

## Introduction

Colorectal cancer is the third most common neoplasm and the third leading cause of cancer death in both men and women across most ethnic-racial groups [1]. Intake of various dietary factors, most notably, meat, fruits/vegetables, and fiber, have been extensively investigated in relation to colorectal cancer risk. Overall, the evidence suggests that consumption of red and processed meat modestly increase the risk of colorectal cancer [2,3]; and fruits [4], vegetables [4,5], and fiber [6–8] decrease risk, although these associations have not been observed across all studies [2,9,10], perhaps due to methodological differences and unaccounted modifying effects.

More recently, studies have focused on the potential modifying effects of common genetic variants, single nucleotide polymorphisms (SNPs), on the relationship between dietary factors and risk of colorectal cancer. However, attention has largely focused on candidate SNPs in genes directly involved in the metabolism of selected nutrients; for example, metabolism of B-vitamins [11], key nutrients found in fruits and vegetables; or the metabolism of carcinogenic by-products resulting from cooking or processing of meat [12]. From these candidate gene/pathway-approaches, few genetic variants have been consistently identified and further investigation is warranted.

Large datasets from genome-wide association studies of colorectal cancer are now available for a comprehensive analysis of gene-diet interactions on the risk of colorectal cancer. To date, one genome-wide study of gene-diet interactions focusing on microsatellite stable/microsatellite-instability low colorectal cancer (1,191 cases, 990 controls) reported no statistically significant gene-diet interactions after replication in an independent dataset [13]. The authors highlighted the need for collaborative consortia to increase sample size, with central quality control procedures and careful standardization and harmonization of definitions and measurements. Hutter et al., using data from the Genetics and Epidemiology of Colorectal Cancer Consortium (GECCO) on 7,106 colorectal cancer cases and 9,723 controls from 9 studies focused on 10 previously identified colorectal cancer-susceptibility loci and conducted a systematic search for interaction with selected lifestyle and dietary factors. The strongest statistical evidence was observed for interaction for vegetable consumption and rs16892766, located on chromosome 8q23.3 near the *EIF3H* and *UTP23* genes ( $p = 1.3E-04$ ) [14].

In this large combined analysis using GECCO from 10 case-control and nested cohort studies comprising 9,287 colorectal cancer cases and 9,120 controls, we build upon these previous reports [13,14] to examine over 2.7 million common polymorphisms for multiplicative interactions with selected dietary factors (red meat, processed meat, fiber, fruit and vegetables) and risk of colorectal cancer. For our primary analyses we used conventional case-control logistic regression that included an interaction term as well as our recently developed Cocktail method, which integrates several novel GxE methods to improve statistical power under various scenarios [15].

## Results

Characteristics of the 10 studies are described in Table S1. Mean intake and quartile cut points of each dietary factor per study are provided in Table S2 and S3. Across all studies we observed an increase in colorectal cancer risk for red meat consumption ( $OR_{\text{per quartile}} = 1.15, p = 1.6E-18$ ) and processed meat consumption ( $OR_{\text{per quartile}} = 1.11, p = 4.2E-09$ ). Decreased colorectal cancer risk was observed for vegetable intake ( $OR_{\text{per quartile}} = 0.93, p = 8.2E-05$ ), fruit intake ( $OR_{\text{per quartile}} = 0.93, p = 1.9E-05$ ) and fiber intake ( $OR_{\text{per quartile}} = 0.91, p = 5.6E-05$ , Figure 1).

Using conventional case-control logistic regression to test for multiplicative interactions we identified a genome-wide significant interaction between variants at chromosome 10p14 and processed meat (Table 1). Within the 10p14 region rs4143094 showed the most significant interaction with processed meat ( $OR_{\text{interaction}}$  for each copy of T-allele and increasing quartile of processed meat = 1.17,  $p = 8.73E-09$ , Table 1 and Figure 2), with no evidence of heterogeneity ( $p_{\text{heterogeneity}} = 0.78$ ). This SNP (rs4143094), as well as correlated SNPs surrounding the rs4143094 SNP, indicate a strong signal peak in the 10p14 region near the *GATA3* gene; as expected SNPs less correlated with rs4143094 show less significant interactions (Figure 3). Stratified by genotype, the risk for colorectal cancer associated with each increasing quartile of processed meat was increased in individuals with the rs4143094-TG and -TT genotypes ( $OR = 1.20$ , 95% CI = 1.13–1.26 and  $OR = 1.39$ , 95% CI = 1.22–1.59, respectively) and null in individuals with the rs4143094-GG genotype ( $OR = 1.03$ , 95% CI = 0.98–1.07, Table 2). Results are very similar for minimal and multivariable adjusted ORs. In addition, the stratified results Table S4 show

## Author Summary

High intake of red and processed meat and low intake of fruits, vegetables and fiber are associated with a higher risk of colorectal cancer. We investigate if the effect of these dietary factors on colorectal cancer risk is modified by common genetic variants across the genome (total of about 2.7 million genetic variants), also known as gene-diet interactions. We included over 9,000 colorectal cancer cases and 9,000 controls that were not diagnosed with colorectal cancer. Our results provide strong evidence for a gene-diet interaction and colorectal cancer risk between a genetic variant (rs4143094) on chromosome 10p14 near the gene *GATA3* and processed meat consumption ( $p=8.7E-09$ ). This genetic locus may have interesting biological significance given its location in the genome. Our results suggest that genetic variants may interact with diet and in combination affect colorectal cancer risk, which may have important implications for personalized cancer care and provide novel insights into prevention strategies.

interaction results using one common reference group. This common SNP (average allele frequency of T allele = 0.25) was directly genotyped in most studies or imputed with high accuracy (imputation  $r^2 > 0.89$ ). With the other dietary factors evaluated, no interactions using the conventional case-control logistic regression analysis reached the genome-wide significance threshold (Table S5).

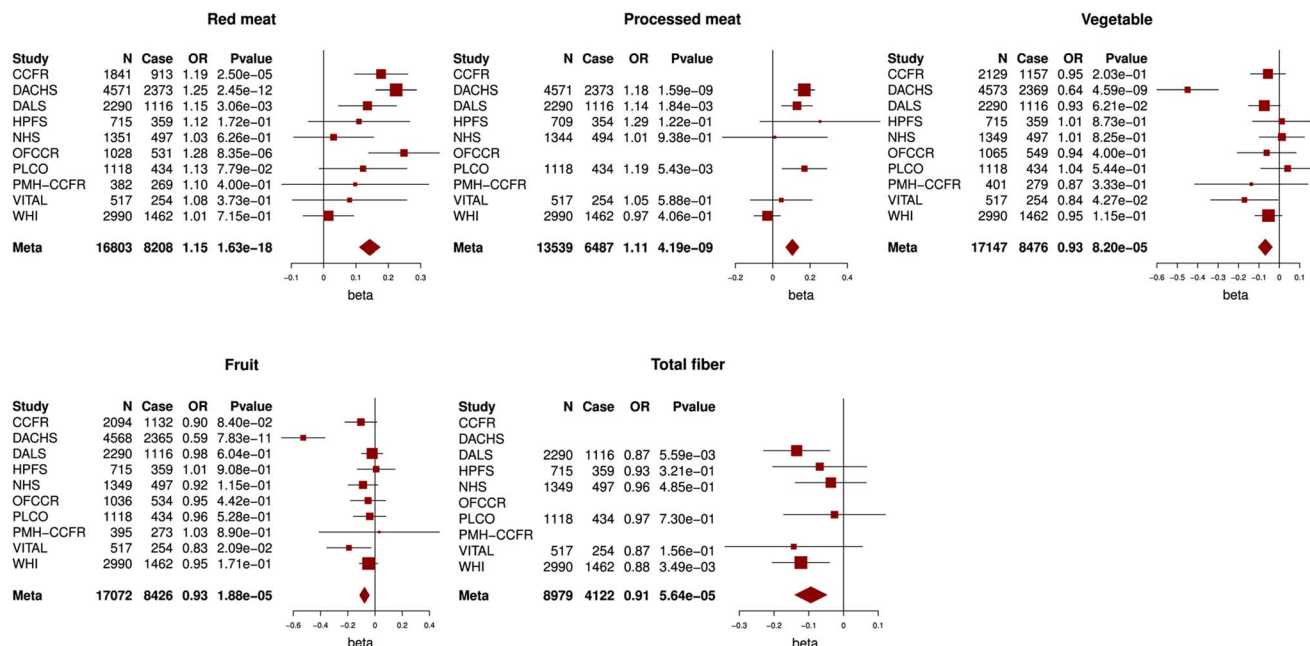
With the other dietary factors, no interactions with any of the 2.7M SNPs were statistically significant using the conventional logistic regression analysis. Furthermore, we did not observe any novel interactions using our Cocktail method or the two exploratory statistical methods by Gauderman et al. [16] and Dai et al. [17] (data not shown).

## Discussion

Genome-wide scans have successfully identified numerous risk loci for colorectal cancer; consortia pooling multiple studies for increased statistical power have continued to identify additional susceptibility loci [18–24]. However, only limited work has been pursued at a genome-wide scale to identify gene-diet interactions. Using individual-level data from ten studies with harmonized dietary intake variables on a total of over 9,000 cases and 9,000 controls, we have conducted a genome-wide analysis for GxE interactions. Using conventional statistical methods, as well as our novel method aiming to improve statistical power, we provide evidence for a novel interaction between rs4143094 and processed meat intake.

The variants in the 10p14 region interacting with processed meat consumption reside within and upstream of *GATA* binding protein 3 (*GATA3*) gene. *GATA3* has long been associated with T cell development, specifically Th2 cell differentiation [25]. *GATA3* is up-regulated in ulcerative colitis [26], which is associated with increased risk of colorectal cancer [27]. However, the role of *GATA* genes as transcription factors extends to epithelial structures with a known role in breast, prostate and other cancers [28–30]. *GATA* factors are involved in cellular maturation with proliferation arrest and cell survival. Loss of *GATA* genes or silencing of expression have been described for breast, colorectal and lung cancers [30].

To further explore this locus, we evaluated the potential functional impact of the most significant SNP in this locus as well as correlated SNPs querying multiple bioinformatics databases, such as Encode and NIH Roadmap (Table S6). The most significant SNP rs4143094 is about 7.2 kb upstream of *GATA* and resides in a 9.5 kb LD block ( $r^2 > 0.8$ ) containing 19 highly correlated SNPs, including rs1269486, which shows the third most significant interaction in this region (Table 1). The rs1269486 variant is located 1420 bases upstream of *GATA3* in a region of



**Figure 1. Associations between red and processed meat, vegetable, fruit and fiber intake and colorectal cancer risk.** Odds ratios (ORs) per quartile of increasing intake, lowest quartile = reference group, N = total number of subjects, case = number of cases. doi:10.1371/journal.pgen.1004228.g001

**Table 1.** Top three SNPs according to lowest p-value for interactions with processed meat for risk of colorectal cancer using conventional case-control logistic regression approach.

SNP	Chr	Position	Context	Gene	CountAllele	CAF*	OR <sub>interaction</sub> **	95% CI	P <sub>interaction</sub>	P <sub>heterogeneity</sub>
rs4143094	10p14	8129142	promoter	GATA3	T	0.21–0.27	1.17	1.11–1.23	8.73E-09	0.78
intergenic										
rs485411	10p14	8133191	promoter	GATA3	C	0.20–0.27	1.18	1.11–1.25	1.72E-08	0.70
non-coding transcript variant										
rs1269486	10p14	8136205	promoter	GATA3	A	0.22–0.26	1.18	1.11–1.25	7.53E-08	0.65

\* CAF, count allele frequency. Count (or tested) allele is defined as the allele that was coded as 1 in the logistic regression (the other allele was coded as 0).

\*\* interaction OR for each copy of the count allele and for each increasing quartile of processed meat intake.  
doi:10.1371/journal.pgen.1004228.t001

open chromatin (DNase I hypersensitivity) with histone methylation patterns consistent with promoter activity in a colorectal cancer cell line (CACO2; Figure S1). As would be expected of a promoter region, experimental evidence supports Pol2 binding along with the transcription factors c-Fos, JunD, and c-Jun [31]. Many of the other SNPs upstream of *GATA3* are located in *GATA3*-antisense RNA1 (*GATA3-AS1*) (formerly FLJ45983). *GATA3-AS1* is a non-coding RNA that may regulate *GATA3* transcript levels in the cell. Further studies are required to elucidate the relationship between *GATA3* and *GATA3-AS1* and determine whether variants in the 10p14 region cause perturbations in regulation.

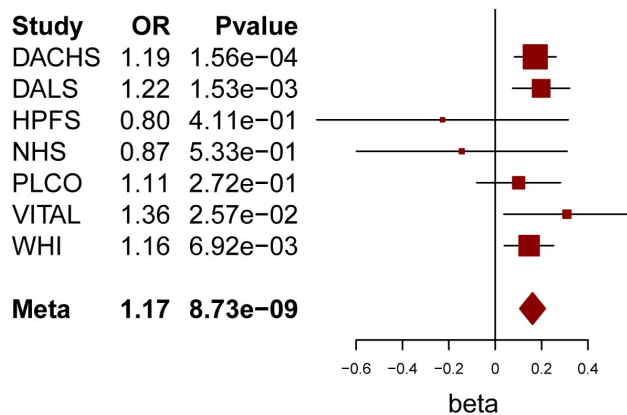
A plausible though speculative biological basis for our findings is that processed meat triggers a pro-tumorigenic inflammatory or immunological response [32] that may necessitate proper *GATA3* transcription levels. Nonetheless, the precise mechanism by which deregulation of *GATA3* is linked to colorectal cancer upon consumption of high levels of processed meat remains unclear. Further study of the role of variants in *GATA3* in colorectal cancer will yield more insight into their functional significance.

The interaction between variants in locus 10p14 and processed meat were identified by the conventional case-control logistic regression analysis. This locus was not identified through our Cocktail method or any of the other exploratory methods (Text S2). However, this is not surprising given that the SNPs in this locus are not strongly associated with colorectal cancer ( $p=0.26$  for rs4143094) and not strongly correlated with processed meat ( $p=0.25$  for rs4143094) and, accordingly, SNPs in this locus were not prioritized in the Cocktail analysis. However, we were somewhat surprised to not identify additional interactions with any of the dietary factors using our Cocktail method, given the expected improvement in power under various scenarios. We recognize that the field of GxE analyses is at an early stage compared with studies for marginal gene-diseases associations. It will be important to see more large-scale empirical GxE studies to judge the impact and potential power gain of the novel GxE methods.

Our analysis has some limitations and notable strengths. We adopted a flexible approach to data harmonization of dietary factors, in a similar fashion to those proposed by other projects [33,34]. We focused on dietary variables that were collected in a similar manner and allowed for harmonization across a large subset of the studies. Ideally, our findings will be replicated in other populations. While a substantial larger number of GWAS have been conducted for colorectal cancer, limited studies have collected information on processed meat and other dietary variables. In the present study, we did not divide our large sample into discovery and replication sets, as it has been shown that the most powerful analytical approach is a combined analysis across all studies [35]. This approach is increasingly used as more samples with GWAS data are becoming available [36]. Importantly, we observed no evidence of heterogeneity in the estimates by study, which suggests that results are consistent across studies.

We not only used the conventional case-control logistic regression, but also took advantage of our recently developed Cocktail method as a second primary analysis approach to potentially improve statistical power. We note that even though for the Cocktail method different interaction tests (case-only and case-control) were used depending on the screening step, the overall genome-wide type I error is controlled at 0.05 (genome-wide level of  $\alpha$  was set to 5E-08), just like the conventional case-control method. As we investigated five dietary factors and used two primary methods additional adjustment for multiple comparisons may be warranted. However, we want to point out that the dietary variables were correlated, e.g. correlation between fruits and





**Figure 2. Forest plot for meta-analysis of interaction analysis for rs4143094 and processed meat.** Odds ratios (ORs) and 95% confidence intervals (95% CI) are presented for each additional copy of the count (or tested) allele (T) and for each increasing quartile of processed meat intake in the multiplicative interaction model. The box sizes are proportional in size to the inverse of the variance for each study, and the lines visually depict the confidence interval. Results from the fixed-effects meta-analysis are shown as diamonds. The width of the diamond represents the confidence interval.  
doi:10.1371/journal.pgen.1004228.g002

vegetables was 0.38, between fruits and fiber was 0.52 or between red and processed meat was 0.62 adjustments for these not independent test is less straight forward. Similarly, the primary methods are not independent from each other, for instance the testing step of the Cocktail method used the case-control or case-only testing, which are consistent or correlated with the conventional case-control analysis. Accordingly, additional multiple comparison adjustment for 5 variables and 2 tests would be too conservative, nevertheless our interaction finding for 10p14 and processed meat would likely remain marginally significant.

With the investment of large GWAS consortium built on well-characterized studies, we are now well-positioned to identify potential interactions between genetic loci and environmental risk factors with respect to colorectal cancer risk. In this study, we have identified a novel interaction between rs4143094 and processed meat. This genetic locus may have interesting biological significance given its proximity to genes plausibly associated with pathways relevant to colorectal carcinogenesis. Nonetheless, further functional analysis is required to uncover the specific mechanisms by which this genetic locus modulates the association between intake of processed meat and colorectal cancer risk.

## Materials and Methods

### Study participants

This analysis uses data from the Colon Cancer Family Registry (CCFR) and the Genetics and Epidemiology of Colorectal Cancer Consortium (GECCO, Text S1 and Table S1) as described previously [14,37]. All cases were defined as colorectal adenocarcinoma and confirmed by medical records, pathologic reports, or death certificate. All studies received ethical approval by their respective Institutional Review Boards and participants gave written informed consent.

### Genotyping, quality assurance/quality control and imputation

Average sample and SNP call rates, and concordance rates for blinded duplicates have been previously published [37]. In brief,

genotyped SNPs were excluded based on call rate ( $<98\%$ ), lack of Hardy-Weinberg Equilibrium in controls (HWE,  $p < 1 \times 10^{-4}$ ), and low minor allele frequency (MAF). We imputed the autosomal SNPs of all studies to the CEU population in HapMap II. SNPs were restricted based on per-study minor allele count  $>5$  and imputation accuracy ( $R^2 > 0.3$ ) to avoid missing any interactions. After imputation and quality control (QC) analyses, approximately 2.7M SNPs were used in the analysis.

All analyses were restricted to individuals of European ancestry, defined as samples clustering with the Utah residents with Northern and Western European ancestry from the CEPH collection (CEU) population in principal component analysis [38], including the HapMap II populations as reference.

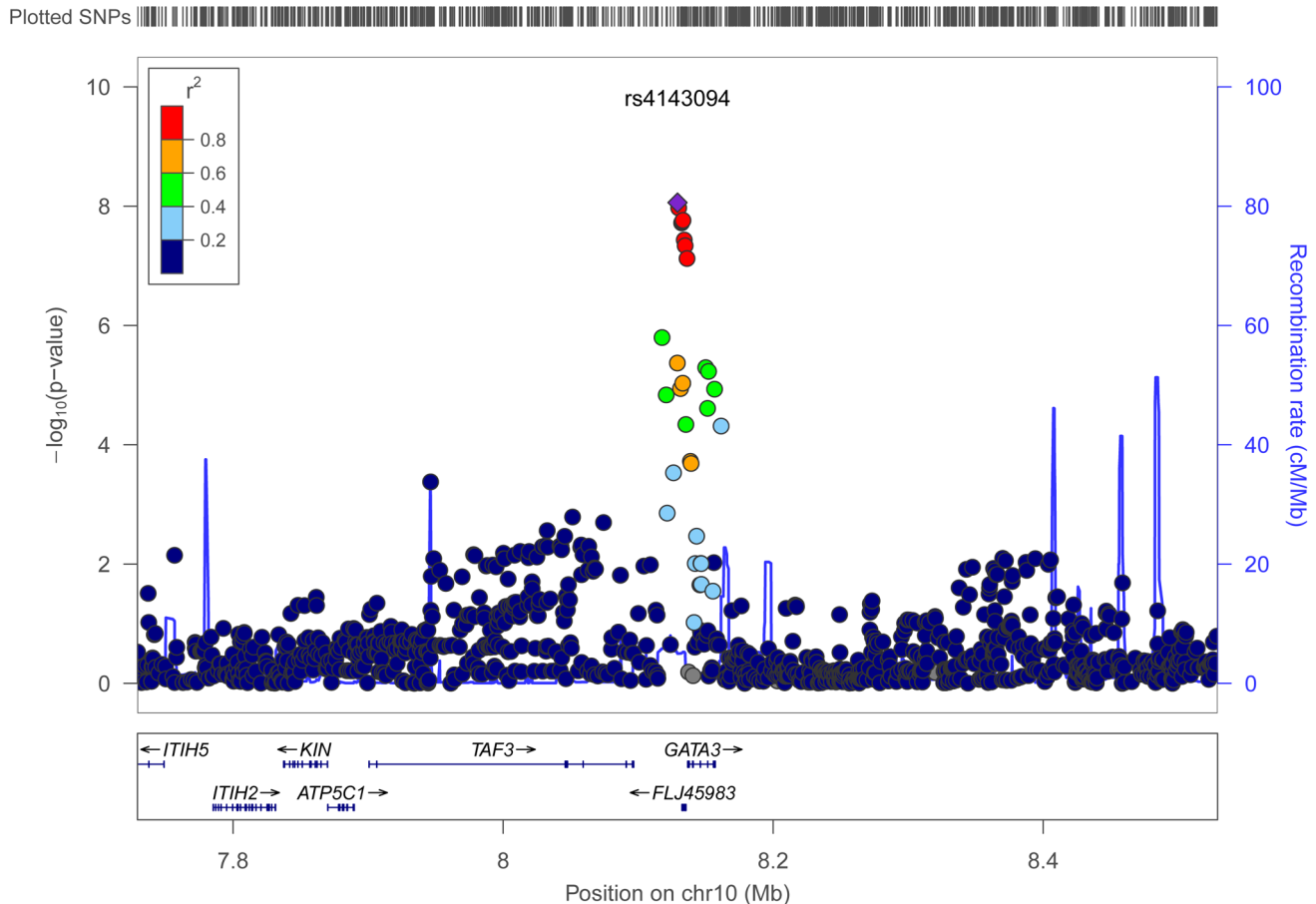
### Harmonization of dietary factors

Information on basic demographics and environmental risk factors was collected by using in-person interviews and/or structured questionnaires, as detailed previously [39–48]. The multi-step data harmonization procedure applied in this study is described in detail by Hutter et al. [14]. Here we focus on selected dietary variables for intake of red and processed meat, fruits, vegetables (all measured in servings per day) and fiber (measured as g/day). These variables were coded as sex- and study-specific quartiles, where the quartile groups were coded 1 to 4 of the quartile within the controls of each study and sex. For studies that due to limited number of questions assessed dietary intake in categories rather than as continuous variables and had less than 4 intake categories, we assigned these categories to the 2<sup>nd</sup> and 3<sup>rd</sup> or 1st to 3<sup>rd</sup> quartile, as appropriate. The lowest category of exposure was used as the reference and each dietary factor was analyzed as an ordinal variable (e.g., 1, 2, 3, 4) in the model. Data harmonization was performed using SAS and T-SQL.

### Statistical methods

Statistical analyses of all samples were conducted centrally at the GECCO coordinating center on individual-level data to ensure a consistent analytical approach. Unless otherwise indicated, we adjusted for age at the reference time, sex (when appropriate), center (when appropriate), total energy consumption (if available) and the first three principal components from EIGENSTRAT to account for potential population substructure. The dietary variables were coded as described above. Each directly genotyped SNP was coded as 0, 1, or 2 copies of the variant allele. For imputed SNPs, we used the expected number of copies of the variant allele (the “dosage”), which has been shown to give unbiased test statistics [49]. Genotypes were treated as continuous variables (i.e. log-additive effects). Each study was analyzed separately using logistic regression models and study-specific results were combined using fixed-effects meta-analysis methods to obtain summary odds ratios (ORs) and 95% confidence intervals (CIs) across studies. We calculated the heterogeneity p-values by Woolf’s test [50]. Quantile-quantile (Q-Q) plots were assessed to determine whether the distribution of the p-values was consistent with the null distribution (except for the extreme tail).

To test for interactions between SNPs and dietary risk factors, we conduct two primary analyses: 1) conventional case-control logistic regression analysis including a multiplicative interaction term; 2) our newly developed Cocktail method [15]. For the conventional logistic regression analysis, we modeled the SNP by environment (GxE) interaction by the product of the SNP and the dietary variable (which is in this study the E), adjusting for age, sex, study site, energy, principal components and the main effects of the SNP and dietary variable. Adjustment for additional variables, smoking, alcohol, BMI and other dietary variables did not



**Figure 3. Regional association results for the interaction between processed meat and rs4143094 with surrounding SNPs.** The top half of the figure has physical position along the x-axis, and the  $-\log_{10}$  of the meta-analysis p-value of the interaction term on the y-axis. Each dot on the plot represents the p-value of the interaction for one SNPxD in relation to colorectal cancer conducted across all studies. The most significant SNP in the region (index SNP) is marked as a purple diamond. The color scheme represents the pairwise correlation ( $r^2$ ) for the SNPs across the region with the index SNP. Correlation was calculated using the HapMap CEU data. The bottom half of the figure shows the position of the genes across the region. These regional association plots are also known as LocusZoom plots.  
doi:10.1371/journal.pgen.1004228.g003

appreciably change the results. A two-sided p-value of  $5 \times 10^{-8}$  for a SNP-diet factor interaction was considered statistically significant, yielding a genome-wide significance level 0.05 assuming about 1 million independent tests across the genome ( $0.05/1,000,000 = 5 \times 10^{-8}$ ) [51–56].

Motivated by recent advances in methods development for detecting GxE interaction [17,57–60], our second approach was based on our recently developed Cocktail method. This statistical method combines the most appealing aspect of several newly developed GxE methods with the goal of creating a comprehensive and powerful test for genome-wide detection of GxE [15]. In brief, this method consists of two-steps: a screening step to prioritize SNPs and a testing step for GxE interaction. Specifically, for the screening step, we ranked and prioritized variants through a genome-wide screen of each of the 2.7M SNPs (referred to as “G”) by the maximum of the test statistics from marginal association of Gs on disease risk [58], and correlation between G and environmental/dietary variable (E) in cases and controls combined [59], a combination which allows for identifying variants with different interaction patterns.

Based on the ranks of these SNPs from screening, we used a weighted hypothesis framework to partition SNPs into groups with higher ranked groups having less stringent alpha-level cut-offs for

interaction [60,61]. We followed the grouping scheme used by Ionita et al. [61] such that for example, the first 3 groups consist of 5 SNPs (SNP 1 to 5), 10 SNPs (SNP 6 to 15) and 20 SNPs (SNP 16 to 36), and the corresponding cut-offs are  $\alpha_{\text{group } 1} = \alpha / (2 \times 5) = 0.005$ ,  $\alpha_{\text{group } 2} = \alpha / (4 \times 10) = 0.00125$  and  $\alpha_{\text{group } 3} = \alpha / (8 \times 20) = 0.0003$ , respectively, so on and so forth, to maintain the overall genome-wide alpha level of 0.05. To avoid testing correlated SNPs, we pruned SNPs based on proximity (exclude any SNP within  $\pm 50$  kb of the selected SNP) given that LD pruning is difficult to implement for large number of SNPs. While the choice of the group size is arbitrary our simulation study showed that different group size did not impact the results substantially, and importantly, we chose the group size before looking at the results.

The second step of the Cocktail method is the testing step. We tested each of the G’s for GxE interactions using the case-only (CO) logistic regression test. The use of the CO test is justified because we did not observe correlation between G and any of the tested dietary factors, and it has been shown that under the independence assumption the CO test provides substantial efficiency gain over the conventional CC test [62]. Since the CO is not independent of the correlation screening (a requirement to avoid inflation of type I error rates) [63], we used CO test only

**Table 2.** Association of processed meat and risk of colorectal cancer by genotype strata for rs4143094.

Adjustment factors	rs4143094	N Case	N Control	Association per quartile of processed meat intake		
				OR	95% CI	P value
Minimal*	GG	3627	3986	1.03	0.98–1.07	0.28
	TG	2428	2610	1.2	1.13–1.26	2.70E-10
	TT	430	445	1.39	1.22–1.59	1.10E-06
Multivariable**	GG	3542	3887	0.98	0.93–1.03	0.5
	TG	2375	2547	1.14	1.08–1.22	1.18E-05
	TT	418	439	1.36	1.18–1.56	1.35E-05

\*Minimal adjusted models included age, sex, study site, energy and PCs.

\*\*Multivariable adjusted models additionally included: BMI, smoking, alcohol and other dietary factors.

Multivariable-adjusted analysis is limited to samples with available data for all covariates used in the analysis.

doi:10.1371/journal.pgen.1004228.t002

when the maximum screening test statistic came from the marginal association, and the case-control test otherwise.

In Text S2, we describe two secondary statistical GxE methods that we used to explore other novel GxE methods: the 2-step method by Gauderman et al. method [16] and a 2 degree of freedom joint test for marginal associations of G and GxE interaction by Dai et al. [17]. All analyses were conducted using the R programming language [64].

## Supporting Information

**Figure S1** Functional annotation of rs4143094 and correlated SNPs in chromosome 10.  
(PDF)

**Table S1** Descriptive characteristics of each study.  
(DOCX)

**Table S2** Mean intake of red meat, processed meat, vegetable, fruit and fiber intake by study.  
(DOCX)

**Table S3** Quartile cut points for intake of red meat, processed meat, vegetable, fruit and fiber intake by study and sex.  
(DOCX)

**Table S4** Interaction between rs4143094 and processed meat intake for risk of colorectal cancer based on one common reference group and stratified analysis by genotype (last row) and by quartiles of processed meat (last column).  
(DOCX)

**Table S5** Top three most significant GxE interactions for red meat, vegetable, fruit and fiber using conventional case-control logistic regression analyses (for regions with multiple highly correlated SNPs only the most significant SNP was included).  
(DOCX)

**Table S6** Description of bioinformatics tools used for functional follow-up of non-coding regions.  
(DOCX)

**Text S1** Study populations. Description of the methodology and individual study populations included in this meta-analysis.  
(DOCX)

**Text S2** Additional statistical analysis. Description of the additional statistical methods used in this meta-analysis.  
(DOCX)

**Text S3** Functional annotation of identified loci. Description of the methodology for functionally annotating significant loci.  
(DOCX)

**Text S4** References supplementary text.  
(DOCX)

## Acknowledgments

**DACHS:** We thank all participants and cooperating clinicians, and Ute Handte-Daub, Renate Hettler-Jensen, Utz Benscheid, Muhabbet Celik and Ursula Eilber for excellent technical assistance.

**GECCO:** The authors would like to thank all those at the GECCO Coordinating Center for helping bring together the data and people that made this project possible. Members of the GECCO consortium are: Brent Zanke, Mathieu Lemire, Steven Gallinger, Thomas Hudson, Roger Green, Sébastien Küry, Stéphane Bezieau, Graham Casey, Loïc Le Marchand, Eric Jacobs, Peter Campbell, Hermann Brenner, Jenny Chang-Claude, Bette Caan, John Potter, Martha Slattery, Flora Qu, Jian Gong, Keith Curtis, Li Hsu, Paul Auer, Riki Peters, Shuo Jiao, Tabitha Harrison, Yi Lin, Andrew Chan, Brian Henderson, Laurence Kolonel, Gad Rennert, Stephen Gruber, Jing Ma, Richard Hayes, Robert Schoen, Stephen Chanock, Polly Newcomb, David Duggan and Emily White.

**HPFS and NHS:** We would like to acknowledge Patrice Soule and Hardeep Ranu of the Dana Farber Harvard Cancer Center High-Throughput Polymorphism Core who assisted in the genotyping for NHS and HPFS, under the supervision of Dr. Immaculata Devivo and Dr. David Hunter, Qin (Carolyn) Guo and Lixue Zhu who assisted in programming for NHS and HPFS. We would like to thank the participants and staff of the Nurses' Health Study and the Health Professionals Follow-Up Study, for their valuable contributions as well as the following state cancer registries for their help: AL, AZ, AR, CA, CO, CT, DE, FL, GA, ID, IL, IN, IA, KY, LA, ME, MD, MA, MI, NE, NH, NJ, NY, NC, ND, OH, OK, OR, PA, RI, SC, TN, TX, VA, WA, WY.

**PLCO:** The authors thank Drs. Christine Berg and Philip Prorok, Division of Cancer Prevention, National Cancer Institute, the Screening Center investigators and staff or the Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial, Mr. Tom Riley and staff, Information Management Services, Inc., Ms. Barbara O'Brien and staff, Westat, Inc., and Drs. Bill Kopp, Wen Shao, and staff, SAIC-Frederick. Most importantly, we acknowledge the study participants for their contributions to making this study possible.

**PMH-CCFR:** The authors would like to thank the study participants and staff of the Hormones and Colon Cancer study.

**WHI:** The authors thank the WHI investigators and staff for their dedication, and the study participants for making the program possible.

## Author Contributions

Conceived and designed the experiments: JAB SIB HB BJC GC ATC JCC SJC DD SG ELG RWH TAH RBH BEH MH JLH LH CMH MAJ LNK



LLM ML PAN UP JDP RES DS MLS SNT CMU BWZ. Performed the experiments: DD. Analyzed the data: LH YL WJG SJ UP. Contributed reagents/materials/analysis tools: SJ CQ KRC. Wrote the paper: JCF UP LH. Collected phenotype data and biological samples and contributed these as investigators for their respective study: SIB HB BJC GC ATC JCC SJC SG ELG RWH RBH BEH MH JLH MAJ LNK LLM PAN JDP RES

DS MLS SNT EW BWZ. Critically reviewed the manuscript drafts and approved the final manuscript: JCF LH CMH YL PTC JAB SIB SJ GC BF ATC MC ML SG TAH LLM PAN MLH BJC CSC BWZ SAR HB ELG KW JCC SJC KRC DD JG RWH RBH MH JLH MAJ LNK CQ AR RES FRS DS DLS SNT MT GSW BEH CMU WJG JDP EW UP.

## References

- Siegel R, Naishadham D, Jemal A (2012) Cancer statistics, 2012. *CA Cancer J Clin* 62: 10–29.
- Alexander DD, Cushing CA (2011) Red meat and colorectal cancer: a critical summary of prospective epidemiologic studies. *Obes Rev* 12: e472–493.
- Alexander DD, Miller AJ, Cushing CA, Lowe KA (2010) Processed meat and colorectal cancer: a quantitative review of prospective epidemiologic studies. *Eur J Cancer Prev* 19: 328–341.
- van Duijnhoven FJ, Bueno-De-Mesquita HB, Ferrari P, Jenab M, Boshuizen HC, et al. (2009) Fruit, vegetables, and colorectal cancer risk: the European Prospective Investigation into Cancer and Nutrition. *The American journal of clinical nutrition* 89: 1441–1452.
- Wu QJ, Yang Y, Vogtmann E, Wang J, Han LH, et al. (2013) Cruciferous vegetables intake and the risk of colorectal cancer: a meta-analysis of observational studies. *Annals of oncology : official journal of the European Society for Medical Oncology/ESMO* 24: 1079–1087.
- Nomura AM, Hankin JH, Henderson BE, Wilkens LR, Murphy SP, et al. (2007) Dietary fiber and colorectal cancer risk: the multiethnic cohort study. *Cancer causes & control : CCC* 18: 753–764.
- Park Y, Hunter DJ, Spiegelman D, Bergkvist L, Berrino F, et al. (2005) Dietary fiber intake and risk of colorectal cancer: a pooled analysis of prospective cohort studies. *JAMA : the journal of the American Medical Association* 294: 2849–2857.
- Dahm CC, Keogh RH, Spencer EA, Greenwood DC, Key TJ, et al. (2010) Dietary fiber and colorectal cancer risk: a nested case-control study using food diaries. *Journal of the National Cancer Institute* 102: 614–626.
- Lin J, Zhang SM, Cook NR, Rexrode KM, Liu S, et al. (2005) Dietary intakes of fruit, vegetables, and fiber, and risk of colorectal cancer in a prospective cohort of women (United States). *Cancer causes & control : CCC* 16: 225–233.
- Ollberding NJ, Wilkens LR, Henderson BE, Kolonel LN, Le Marchand L (2012) Meat consumption, heterocyclic amines and colorectal cancer risk: the Multiethnic Cohort Study. *International journal of cancer Journal international du cancer* 131: E1125–1133.
- Liu AY, Scherer D, Poole E, Potter JD, Curtin K, et al. (2013) Gene-diet-interactions in folate-mediated one-carbon metabolism modify colon cancer risk. *Molecular nutrition & food research* 57: 721–734.
- Cotterchio M, Boucher BA, Manno M, Gallinger S, Okey AB, et al. (2008) Red meat intake, doneness, polymorphisms in genes that encode carcinogen-metabolizing enzymes, and colorectal cancer risk. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology* 17: 3098–3107.
- Figueiredo JC, Lewinger JP, Song C, Campbell PT, Conti DV, et al. (2011) Genotype-environment interactions in microsatellite stable/microsatellite instability-low colorectal cancer: results from a genome-wide association study. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology* 20: 758–766.
- Hutter CM, Chang-Claude J, Slattery ML, Pflugeisen BM, Lin Y, et al. (2012) Characterization of gene-environment interactions for colorectal cancer susceptibility loci. *Cancer research* 72: 2036–2044.
- Hsu L, Jiao S, Dai JY, Hutter C, Peters U, et al. (2012) Powerful cocktail methods for detecting genome-wide gene-environment interaction. *Genetic epidemiology* 36: 183–194.
- Gauderman WJ, Zhang P, Morrison JL, Lewinger JP (2013) Finding novel genes by testing G×E interactions in a genome-wide association study. *Genetic epidemiology* 37: 603–613.
- Dai JY, Kooperberg C, Leblanc M, Prentice RL (2012) Two-stage testing procedures with independent filtering for genome-wide gene-environment interaction. *Biometrika* 99: 929–944.
- Tenesa A, Farrington SM, Prendergast JG, Porteous ME, Walker M, et al. (2008) Genome-wide association scan identifies a colorectal cancer susceptibility locus on 11q23 and replicates risk loci at 8q24 and 18q21. *Nat Genet* 40: 631–637.
- Tomlinson IP, Webb E, Carvajal-Carmona L, Broderick P, Howarth K, et al. (2008) A genome-wide association study identifies colorectal cancer susceptibility loci on chromosomes 10p14 and 8q23.3. *Nat Genet* 40: 623–630.
- Broderick P, Carvajal-Carmona L, Pittman AM, Webb E, Howarth K, et al. (2007) A genome-wide association study shows that common alleles of SMAD7 influence colorectal cancer risk. *Nat Genet* 39: 1315–1317.
- Tomlinson I, Webb E, Carvajal-Carmona L, Broderick P, Kemp Z, et al. (2007) A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. *Nat Genet* 39: 984–988.
- Zanke BW, Greenwood CM, Rangrej J, Kustra R, Tenesa A, et al. (2007) Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. *Nat Genet* 39: 989–994.
- Houlston RS, Webb E, Broderick P, Pittman AM, Di Bernardo MC, et al. (2008) Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. *Nat Genet* 40: 1426–1435.
- Jia WH, Zhang B, Matsuo K, Shin A, Xiang YB, et al. (2012) Genome-wide association analyses in east Asians identify new susceptibility loci for colorectal cancer. *Nature genetics* 45: 191–196.
- Hosoya T, Maillard I, Engel JD (2010) From the cradle to the grave: activities of GATA-3 throughout T-cell development and differentiation. *Immunol Rev* 238: 110–125.
- Christophi GP, Rong R, Holtzapfel PG, Massa PT, Landas SK (2012) Immune markers and differential signaling networks in ulcerative colitis and Crohn's disease. *Inflammatory bowel diseases* 18: 2342–2356.
- Gupta RB, Harpaz N, Itzkowitz S, Hossain S, Matula S, et al. (2007) Histologic inflammation is a risk factor for progression to colorectal neoplasia in ulcerative colitis: a cohort study. *Gastroenterology* 133: 1099–1105; quiz 1340–1091.
- Chou J, Provot S, Werb Z (2010) GATA3 in development and cancer differentiation: cells GATA have it! *Journal of cellular physiology* 222: 42–49.
- Nguyen AH, Tremblay M, Haigh K, Koumakpayi IH, Paquet M, et al. (2013) Gata3 antagonizes cancer progression in Pten-deficient prostates. *Human molecular genetics* 22: 2400–2410.
- Zheng R, Blobel GA (2010) GATA Transcription Factors and Cancer. *Genes Cancer* 1: 1178–1188.
- Rosenbloom KR, Sloan CA, Malladi VS, Dreszer TR, Learned K, et al. (2013) ENCODE data in the UCSC Genome Browser: year 5 update. *Nucleic acids research* 41: D56–63.
- Hedlund M, Padler-Karavani V, Varki NM, Varki A (2008) Evidence for a human-specific mechanism for diet and antibody-mediated inflammation in carcinoma progression. *Proceedings of the National Academy of Sciences of the United States of America* 105: 18936–18941.
- Bennett SN, Caporaso N, Fitzpatrick AL, Agrawal A, Barnes K, et al. (2011) Phenotype harmonization and cross-study collaboration in GWAS consortia: the GENEVA experience. *Genetic epidemiology* 35: 159–173.
- Fortier I, Doiron D, Burton P, Raina P (2011) Invited commentary: consolidating data harmonization—how to obtain quality and applicability? *American journal of epidemiology* 174: 261–264; author reply 265–266.
- Skol AD, Scott LJ, Abecasis GR, Boehnke M (2006) Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nature genetics* 38: 209–213.
- Pearce CL, Rossing MA, Lee AW, Ness RB, Webb PM, et al. (2013) Combined and interactive effects of environmental and GWAS-identified risk factors in ovarian cancer. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology* 22: 880–890.
- Peters U, Jiao S, Schumacher FR, Hutter CM, Aragaki AK, et al. (2013) Identification of Genetic Susceptibility Loci for Colorectal Tumors in a Genome-Wide Meta-analysis. *Gastroenterology* 144: 799–807 e724.
- Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 38: 904–909.
- Newcomb PA, Baron J, Cotterchio M, Gallinger S, Grove J, et al. (2007) Colon Cancer Family Registry: an international resource for studies of the genetic epidemiology of colon cancer. *Cancer Epidemiol Biomarkers Prev* 16: 2331–2343.
- Slattery ML, Potter J, Caan B, Edwards S, Coates A, et al. (1997) Energy balance and colon cancer—beyond physical activity. *Cancer research* 57: 75–80.
- Christen WG, Gaziano JM, Hennekens CH (2000) Design of Physicians' Health Study II—a randomized trial of beta-carotene, vitamins E and C, and multivitamins, in prevention of cancer, cardiovascular disease, and eye disease, and review of results of completed trials. *Annals of epidemiology* 10: 125–134.
- Prorok PC, Andriole GL, Bresalier RS, Buys SS, Chia D, et al. (2000) Design of the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial. *Controlled clinical trials* 21: 273S–309S.
- (1998) Design of the Women's Health Initiative clinical trial and observational study. The Women's Health Initiative Study Group. *Controlled clinical trials* 19: 61–109.
- Hoffmeister M, Raum E, Krtischil A, Chang-Claude J, Brenner H (2009) No evidence for variation in colorectal cancer risk associated with different types of postmenopausal hormone therapy. *Clinical pharmacology and therapeutics* 86: 416–424.

45. Brenner H, Chang-Claude J, Seiler CM, Rickert A, Hoffmeister M (2011) Protection from colorectal cancer after colonoscopy: a population-based, case-control study. *Annals of internal medicine* 154: 22–30.
46. Kury S, Buecher B, Robiou-du-Pont S, Scoul C, Sebille V, et al. (2007) Combinations of cytochrome P450 gene polymorphisms enhancing the risk for sporadic colorectal cancer related to red meat consumption. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology* 16: 1460–1467.
47. Colditz GA, Hankinson SE (2005) The Nurses' Health Study: lifestyle and health among women. *Nature reviews Cancer* 5: 388–396.
48. Giovannucci E, Rimm EB, Stampfer MJ, Colditz GA, Ascherio A, et al. (1994) Aspirin use and the risk for colorectal cancer and adenoma in male health professionals. *Annals of internal medicine* 121: 241–246.
49. Jiao S, Hsu L, Hutter CM, Peters U (2011) The use of imputed values in the meta-analysis of genome-wide association studies. *Genetic epidemiology* 35: 597–605.
50. Woolf B (1955) On estimating the relation between blood group and disease. *Ann Hum Genet* 19: 251–253.
51. Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. *Science* 273: 1516–1517.
52. (2005) A haplotype map of the human genome. *Nature* 437: 1299–1320.
53. (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447: 661–678.
54. Hoggart CJ, Clark TG, De Iorio M, Whittaker JC, Balding DJ (2008) Genome-wide significance for dense SNP and resequencing data. *Genetic epidemiology* 32: 179–185.
55. Pe'er I, Yelensky R, Altshuler D, Daly MJ (2008) Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genetic epidemiology* 32: 381–385.
56. Dudbridge F, Gusnanto A (2008) Estimation of significance thresholds for genomewide association scans. *Genetic epidemiology* 32: 227–234.
57. Mukherjee B, Chatterjee N (2008) Exploiting gene-environment independence for analysis of case-control studies: an empirical Bayes-type shrinkage estimator to trade-off between bias and efficiency. *Biometrics* 64: 685–694.
58. Kooperberg C, Leblanc M (2008) Increasing the power of identifying gene×gene interactions in genome-wide association studies. *Genetic epidemiology* 32: 255–263.
59. Murcray CE, Lewinger JP, Gauderman WJ (2009) Gene-environment interaction in genome-wide association studies. *Am J Epidemiol* 169: 219–226.
60. Roeder K, Wasserman L (2009) Genome-Wide Significance Levels and Weighted Hypothesis Testing. *Statistical science : a review journal of the Institute of Mathematical Statistics* 24: 398–413.
61. Ionita-Laza I, McQueen MB, Laird NM, Lange C (2007) Genomewide weighted hypothesis testing in family-based association studies, with an application to a 100K scan. *American journal of human genetics* 81: 607–614.
62. Piegorsch WW, Weinberg CR, Taylor JA (1994) Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. *Statistics in medicine* 13: 153–162.
63. Dai JY, Kooperberg C, Leblanc M (submitted) On two-stage hypothesis testing procedures via asymptotically independent statistics. *J R Stat Soc Series B Stat Methodol*.
64. (2010) R Development Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria.